

DOCUMENT RESUME

ED 425 189

TM 029 253

AUTHOR Zopp, William B.
TITLE A Study of Comparing Testing Styles and Retention of Material.
PUB DATE 1998-08-00
NOTE 37p.; Master's Thesis, Salem-Teikyo University.
PUB TYPE Dissertations/Theses - Masters Theses (042)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Academic Achievement; Biology; Comparative Analysis; *Essay Tests; *High School Students; High Schools; *Multiple Choice Tests; *Retention (Psychology); *Test Format; Test Use

ABSTRACT

The purpose of this study was to compare two different testing styles to see how they relate to retention of information. The study was conducted during the first semester of the 1996-97 school year at Greenbriar East High School in Lewisburg, West Virginia. The students in the study were in two Biology II classes, one with an enrollment of 26 students and the other with an enrollment of 24. The strategy for the study was: (1) to divide each class into equal numbers, giving half multiple choice tests for 4.5 weeks and the other half essay tests on the same material; and (2) at the end of 4.5 weeks, to switch the type of test each student was taking so that every student was involved in both testing styles. Both groups were taught at the same pace and were exposed to the same material. Students' grades for the previous 9 weeks were used as pretest scores. At the conclusion of the study, the grades of the two testing styles were compared by the use of a two sample t-test. At the 0.05 level of significance, there was a significant difference in the achievement levels of the two groups. Students taking essay tests seemed to do better than those taking multiple choice tests. (Contains 2 tables and 29 references.) (Author/SLD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

TM

ED 425 189

**A Study of Comparing Testing Styles
and Retention of Material**

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL IN MICROFICHE
AND IN ELECTRONIC MEDIA FOR ERIC
COLLECTION SUBSCRIBERS ONLY
HAS BEEN GRANTED BY

William Zopp

2A TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to
improve reproduction quality.

• Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy.

A Thesis

Presented to

**The Faculty of the Graduate School
of Salem-Teikyo University**

**In Fulfillment of the
Requirements for the Degree of
Master of Arts in Education**

By

William B. Zopp

August 1998

TM029253

Salem-Teikyo University
Salem, West Virginia

This thesis submitted by William B. Zopp has been approved meeting the research requirements for the Master of Arts Degree.

8/17/98
Date

Gaby van der Giessen
Thesis Committee Chair
Gaby van der Giessen, Ph. D., Administrator for Graduate Education

9/3/98
Date

Gary S. McAllister
Gary McAllister, Professor of Education

9/8/98
Date

Harry C. Hartman
Harry Hartman, Professor of Physical Education

Abstract

The purpose of this study was to compare two different testing styles and see how they relate to retention of information. The study was conducted during the first semester of the 1996-1997 school year at Greenbrier East High School in Lewisburg West Virginia. The students used in the study comprised two Biology II classes. One class had an enrollment of 26 students and the other had an enrollment of 24. The strategy for the study was:

1. Divide each class in equal numbers. Give half of them multiple choice tests For 4 1/2 weeks and the other half essay tests on the same material.
2. At the end of 4 1/2 weeks, (1/2 the grading period), the type of test that each student was taking was switched so that every student was involved in both testing styles.

Both groups were taught at the same pace and were exposed to the same material. The students' previous nine weeks grade was used as a pretest score. This grading period consisted of a combination of both testing styles. At the conclusion of the study, the grades of the two testing styles were compared by the use of a two sample t-test. At the .05 level of significance there was a significant difference in the achievement levels of the two groups. The study showed that students taking essay tests seemed to do better than those taking multiple choice tests.

Table of Contents

Chapter	Page
1. Purpose of the study	1
Introduction	1
Hypothesis	2
Definition of Terms	2
Limitations	2
Assumptions	3
2. Review of the Literature	4
Pluses of Essay Testing	10
Minuses of Essay Testing	12
Suggestions from the Pluses and Minuses	15
Concern About Traditional Testing	17
3. Description of Population	22
Hypothesis	23
Procedure	23
Research Question	24
4. Statistical Analysis	25
Descriptive Statistics	25
t-test	26

5. Summary, Results and Recommendations	27
Summary	27
Results	27
Recommendations	28
Reference Page	29

CHAPTER ONE

Introduction/Purpose of the Study

The purpose of the study was to compare multiple choice and essay testing to see which method has the greatest effect on the retention of material. This is a topic that has a lot of pros and cons on both methods. The results of this study could benefit a teacher to be a good student evaluator, and test-maker. Most teachers include both of these testing styles along with some tests that were a combination of both (Tuckman, 1993.)

Introduction

It seems that the multiple choice test generally is harder to construct but easier to grade. The grading part is especially true with machines such as the scantron in which each student fills out the answers on a form by coloring in bubbles. The form is then fed into the scantron producing a graded answer form in a few seconds. On the other hand, the essay test could be considered to be just the opposite. Usually they are relatively easy to construct but hard to grade, or at least hard to grade accurately. More often than not the time factor in grading plays a role in which type of test will be chosen by the classroom teacher. Multiple choice tests usually can be graded fairly quickly, while essay tests can take several hours to grade just one class (Murchan, 1989). This study will discuss these two testing methods in much more detail, expanding on the positive and negative aspects of both, while trying to evaluate the amount of material that is retained by the students for a certain time period after the test is administered.

The reason for this study is to help teachers realize which of these testing methods is a more accurate representation of what the student has actually learned, and what they remember. A teacher's grading procedure can be extremely important. Sometimes grades

determine if a student can be admitted to a higher level course, admitted to a particular college, or even get a certain job (Bailey, 1983). Even the difference between a low A and a high B can be significant.

Hypothesis

H₀: Students will achieve at the same level regardless of the style of test they are given.

H₁: When the students are taking the essay tests, they will achieve at higher levels than when they are taking multiple-choice tests.

Definition of Terms

1. Scantron - A machine that can be used to grade a particular type of answer sheet.
2. Formulaic - A prescribed or set form or method of doing something (Webster, 1984).
3. Portfolio assessment - An alternative way of evaluation student progress which involves the student compiling a file on the topic which might include such things as articles, pamphlets, etc (Tuckman, 1991).
4. Scrawl - To write or draw poorly or carelessly (Webster, 1984).
5. Variance - Difference (Webster, 1984).

Limitations

1. The tests are teacher constructed tests.
2. The sample size used was small.
3. The validity of the results will depend very much on the teacher's own ability to produce good tests.
4. Much of the literature used did not compare the retention levels of the students tested.

Assumptions

1. The objectives were tested.
2. The sample used was homogeneous, therefore results of the study were valid.
3. The sample size was adequate.
4. The tests used in the study were good indicators of students retention levels.

CHAPTER TWO

Review of the Literature

One of the most modern day trends in educational evaluation in the United States is the move away from traditional testing methods towards "authentic assessments", which are designed to measure student performance of skills, abilities, and knowledge directly (Jones, 1994).

While there is no consensus as to precisely what constitutes authentic assessment, there is general agreement that it incorporates: (1) emphasis on examinee performance, assessing not only what the examinee knows, but what the examinee can do; (2) use of direct methods of assessment; (3) inclusion of a high degree of realism; and (4) activities for which there may be no one correct answer, in a simulation of realism (Jones, 1994). The primary distinction between traditional testing methods and authentic assessment is the choice of question format (Jones, 1994). Most of the studies done within the educational community suggest a move away from the multiple-choice format to formats that would be considered to reflect higher-order cognitive processes more accurately. In other words, the tests should measure more of what the student actually knows. The results of these studies have caused many teachers to change their testing methods to include alternatives to multiple choice testing. One of the big alternatives seems to be the inclusion of essay questions in their format (Jones, 1994).

One such study comparing multiple choice versus essay testing formats was performed using students from the University of Minnesota and the University of Wisconsin at River Falls. (Thiede, 1991). In this study the formats were designed to test recognition and recall. Results indicated a higher correlation between the essay test format and the actual

recognition and recall of students. An explanation for this is that examinees can best put down on paper what they actually know better than they can choose between several answers on a typical multiple choice test. Of course one logical question for someone looking at these results would be: were these tests both reliable in relation to what was taught by the instructor? This opens the door for a widely discussed topic in the educational community, the validity and reliability of tests (Thiede, 1991).

Teachers often struggle with the construction part of testing. Does the test accurately cover what the students should have learned? Is this test an accurate evaluation tool (Thiede, 1991)?

Over the past few years, State Education Departments nationwide have been examining the role of direct measures of writing in their assessment programs. Lately the emphasis has been switching somewhat to examining how essays can be used to determine student achievement in a wide variety of subject areas (Thiede, 1991). A common conception seems to be that though rater reliability is undeniably not as high as that obtained by objective tests of achievement, the benefit to be gained in terms of measuring higher cognitive skills makes the practice worthwhile (Thiede, 1991 p. 34).

Teachers who give totally multiple choice tests spend a lot less time grading than those teachers who give essay tests, but which teacher is creating a better evaluation of what the student actually knows (Thiede, 1991 p. 35)?

According to Murchan (1989) scores on essay exams were less reliable and susceptible to a significantly larger standard error of measurement than an objective test of equal length. His study dealt with reliability, content validity, and construct validity of the two test formats. The two tests measured different skills, both of them assessing higher and

lower order abilities. There was insufficient evidence to indicate that writing ability or quality of handwriting influenced scores assigned to essays.

Therefore, one would have to assume that the tests were graded as fairly as possible.

As a second part to his study, Murchan studied tests that were a combination of the two formats mentioned. Based on these results, Murchan recommends that all classroom teachers, when trying to construct a "good" test, use a combination of multiple choice questions and essay questions.

A paper written by Edward White, "An Apologia for the Timed Impromptu Essay Test," brings out several strong points against essay testing. White states that twenty years ago the essay test was very well thought of by English departments in colleges all over the country. White, a professor of English at the Cal State University has written several books on the topic of testing formats. His study says that today, essay testing is under attack by many as being formulaic, unresponsive to the nature of writing, and destructive to the curriculum. "A single sitting, essay exam neither tests the skills these students have been taught nor supports the connections between learning and writing." (p. 30)

Peter Elbow, a gentleman who helped in the study says "essay tests give us agreement about a faint, smudged, and distorted picture of the student's writing ability." (White, 1995, p. 30)

White used a survey, (Murphy, et al. p. 25), which stated that over 70% of colleges and universities use some sort of essay tests as part of their assessments. This could range from placement of entering students to certification of departing seniors. For years the faculty at these institutions must have felt comfortable about how they were grading, but

now they find themselves feeling guilty and wondering if they have been testing the most accurate way. Some are moving hesitantly into portfolio assessment, but for many this type of grading is not good because of the substantial investment of time and energy. This also holds true for essay testing. So, is it worth the effort to maintain the essay test, with its limitations and faults, when cheap and easy multiple choice tests are readily available? This is a question that surely goes through the mind of almost every classroom teacher.

White, being an English professor, brings out the point that the majority of educators who use the essay test as a large part of their scoring system are in English. He says that, for almost all other disciplines outside of the English department including administrators, assessment means multiple choice testing. "Evaluation of actual writing, whether on essay tests, term papers, or portfolios, is still generally seen as hopelessly subjective, unreliable, and arbitrary" (White, 1995, p. 33).

Today the supposed efficiency, objectivity, and economy of multiple choice testing remain in the background of most discussions of assessment. Multiple choice tests are efficient because they can be graded so fast and the students receive feedback quickly. It seems that even some of the "better" teachers sometimes take much too long in returning their graded tests.

The most recent survey of testing practices shows that approximately 48% are multiple choice, 49% are essay, and 3% are portfolios. (Murphy et al. p. 22.) This study suggests that the correct form of evaluation is still very much up in the air. Murphy went on to state that the format is ultimately up to the classroom teacher or professor. He hints that it can even be decided because of the teacher's personality or lifestyle. Because a student's grades can greatly affect their future it is important for a teacher to consider this

fact each time an assessment is to be given. If teachers would remember this, grades would probably be much more reliable. But, as educators well know, especially at the pre-college level, time is such a factor that it weighs heavily on the decision of testing format.

Any assessment is, in one sense, a means of gathering information. Thus it should be clear what information is needed and how it will be used, before it is decided on how to gather it. Another way to conceptualize the issue is to think of an assessment as a way of providing answers to questions. The questions need to be well formulated before the answers are sought. These ways of thinking about any assessment seem obvious enough, but very few assessments of any kind actually follow them. In most cases, the method of testing is the first issue decided instead of the last; the answer is sought before the question has become clear. Although this procedure seems obviously backwards, it is practiced quite often.

Dr. White makes his view of the situation very simple when he says, "No assessment device is good or bad in itself but only in context. Only when we know what we are seeking to discover can we claim that a particular kind of assessment is appropriate or not." (White, p. 34) Dr. White's study indicates that multiple choice testing create a situation in which students have to pick a single right answer embedded in a series of wrong answers. Since educators are ultimately trying to prepare students for real life situations and how to deal with them, this type of testing is creating an artificial "real life" situation. Outside the testing room this situation doesn't exist. Furthermore, the right answer is only "right" because the creator of the test has declared it to be so by a certain set of rules. Those who have learned to succeed on such tests do so not by asking which answer is correct in the world or under certain circumstances, but by choosing the one the

test makers are likely to have chosen to fit the needs of the test. "The multiple choice test thus examines, along with its content, the degree to which the student can adapt reality to the needs of authority." (White, p. 34) This indeed may be the reason that many such scores correlate well with success in college. "The required submission to the world of that kind of test may also suggest reasons why minority groups score less well on these so called objective tests than they do on writing tests. (White, 1995, p. 37)

But, the only way to answer an essay test is to write - an obvious fact not mentioned much these days. Writing well on an essay test means recalling information, selecting an appropriate vocabulary, constructing sentences and paragraphs, and somehow, having something to say. This process seems to involve more of the critical thinking skills than simply choosing one correct answer from four or five. Every essay test shares the artificiality of all tests, but it does require an active response rather than the passive submission called for by multiple-choice examinations. While it is naive to imagine an essay test as a totally valid measure of a student, it is also naive to ignore the fact that the production of writing out answers is a wholly different activity than filling in the bubbles on an answer sheet (Murchan, 1989 p.62).

Bruce Tuckman, a professor of educational research at Florida State University, did a study comparing the advantages and disadvantages of the essay test. He summed up his paper with the following quote: "Essay tests, at best, are easily constructed, relatively valid tests of higher cognitive processes; but they are hard to score reliably. They can be improved by using objectives, scoring guides, and other test construction and scoring aides" (Tuckman, 1991 p. 300).

Pluses of Essay Testing

1. Essay tests are easy to construct-or so it would appear.

"Thus, the classroom teacher faced with the alternative of using essay questions, which he finds easy to prepare and not too demanding to read, or objective questions, which he finds difficult to prepare and essentially superficial in what they appear to measure, decides to use essay questions (Coffman, 1971 p.274).

Given the fact that teachers are not particularly well trained in test construction and recognize their own deficiencies (Carter, 1984), the likelihood that they will be able to write good short answer items and refine them based on item analysis is small. While a good solution to this problem would be, as Hill (1991) suggests, giving teachers more training in test construction, this is not likely to happen. Hence, we have Coffman's (1971) conclusion that the essay test is for many teachers the path of least resistance in terms of the effort and skill it's construction requires.

But while constructing an essay test may be considerably less laborious than constructing a multiple-choice test, there are skill requirements according to Soloman (1965). A teacher cannot necessarily just sit down and, in a matter of minutes, produce a good essay test. Teachers must be taught how to fit items to well-prepared objectives, how to focus their items, and how to develop their criteria in order to produce good essay items. How many teachers actually receive this type of training (Green, 1979)?

2. Essay items can be used to measure the higher cognitive processes. Application, analysis, synthesis, and evaluation involve a process as well as a product, elucidation as well as description. It is difficult to measure these processes with short answer items, such

as multiple-choice questions, which limit the action required of the student to naming or identifying. Actions like describing, explaining, demonstrating, and constructing require a less structured format, like that provided by essay items. Again, however, this does not mean that all essay items succeed in requiring higher level thought. Many poor ones measure retentive knowledge, and could well be replaced by short answer items (Ashburn, 1938).

3. Essay items are valid. This, of course, would depend on the nature of the objectives they are constructed to measure. However, they do possess more free validity, in that the performance they require can be considered more realistic or "authentic" than that required by multiple-choice items. By virtue of the fact that essay items yield a sample of student performance rather than a mere word or check mark, teachers have more information to go on in judging a student's grasp of the content. Hence, the validity of essay tests, at least well-written ones, is inherent in the correlation between the requirement to demonstrate a thought process and the evidence of that process having been performed in the written answer (Dietrich 1967).

4. Essay items have positive side effects. In order to answer a well-constructed essay item, a student must think, organize, and write. To answer a well-constructed multiple-choice item, a student at most must only think. Hence, organizing and writing become important and necessary skills in order to produce high-quality essay answers (Dietrich 1967).

Minuses of Essay Testing

1. Essay test responses take considerable time and effort to score. This is an obvious disadvantage to anyone who has ever scored an essay test. Responses are often lengthy, disorganized, and written in a scrawl (Coffman, 1968). If a teacher with five classes of 30 students each gives a test with three essay items, each of which yields about a one page answer, the result is 450 handwritten sheets to be read and evaluated. It would probably take less time and effort, at least for a trained person to write 50 multiple-choice items than it would to read and evaluate 450 handwritten sheets. Since good instructional practice requires test results to be returned soon after they are taken, our fictitious teacher will spend some long evenings reading and scoring (Tuckman, 1993).

Add to the reading and evaluation task that of supplying written feedback and the job becomes even more difficult. Whereas, on a multiple-choice item the result is either right or wrong, individual essay responses have their strengths and weaknesses, and the good teacher is obligated to point them out in writing. The difference between multiple-choice testing and essay testing in terms of total labor turns out not to be that great, except that for multiple-choice testing, the work is in the writing, while for essay testing, the work is in the scoring (Murchan, 1989).

2. It is especially difficult to score essay test responses reliably.

"A significant portion of the variance of systematic or variable error unrelated to the underlying ability one wishes to assess" (Coffman, 1971 p. 276).

It has been found that different scorers often assign different grades to the same essay response; the same scorer often assigns different grades to the same essay response on

different occasions; and these grading differences tend to increase as the essay question permits greater freedom of response. (Noyes, 1963; Vernon and Millican, 1954).

Why is it that differences between scorers tend to occur? One reason is that scorers differ in their severity, some assigning relatively high grades, others relatively low ones. (Coffman and Curfman, 1968). Another difference in scores is that some do not tend to give extreme scores while others do. Also, different scorers do not always use the same criteria for judging the merit of an essay response. Some scorers require that certain points be mentioned while others do not; some deduct points for spelling, sentence structure, and grammar, while others disregard these areas (Coffman and Curfman, 1968).

Diederich (1967) identified five types of responses that scorers reacted to differently in terms of points awarded: ideas, form, flavor, mechanics, and wording. The score assigned to a given response would depend on which of these the scorer thought to be important. But, even a single scorer is inconsistent in judgment, often as a function of fatigue or of bias based on familiarity with other work by the same student. Hughes and Keeling, (1984), found that essay ratings went up or down depending on the quality of the essays that were read immediately before and after a given essay. These represent context errors.

Most of the recent research on essay scoring has been done in conjunction with large-scale testing of writing, where a number of trained judges read every writing sample and scored them on the same scale. In such cases, ratings can be expected to be more consistent because of the training each judge has received. However, teachers do not receive this training-even at the college level (Carter, 1984).

In an old study, Ashburn, (1938), compared the ratings of three professors on two sets of essays and found that the level of scoring agreement did not exceed 50 percent. In fact, 40 percent of the students' scores varied from passing to failing or vice versa, depending on which professor did the scoring. Clearly, the biggest problem with essay testing lies in the realm of scoring. If the retention of certain objectives is what the teacher is evaluating, then the amount of actual retention could be misinterpreted, based on the caliber of the scorer.

2. Some variability in student response is based on factors other than whether or not they have the prerequisite knowledge and understanding. Writing ability, for example, affects the quality of an essay response, and its importance may depend on the complexity of the question (Carter, 1984). Some students are slower at organizing their thoughts, and this quality will be penalized, especially on a complexly worded question. While the writing and organizing requirement of essay items was cited earlier as an advantage in that it would reinforce the importance of these activities in the curriculum, it can also be cited as a disadvantage from a measurement perspective. It introduces variability that may be seen as irrelevant to the knowledge and understanding that is being tested. The broader and more wide-ranging the question, the more these "irrelevant" factors will affect the quality of the response. (Ebel, 1965).

Suggestions from the Pluses and Minuses

1. Start with clearly defined objectives and write essay questions that are well focused (Jones, 1994). Essay questions that are short and straightforward can be scored more consistently than ones that are more complex can. Also, by using shorter questions, it becomes possible to include more of them on a single test. A number of short questions provide a better sample of student performance than a single, complicated one. Multi-part questions can unjustly penalize the test-taker who does not understand the basic premise on which the question is based (Tuckman, 1993). Keep in mind, however, Coffman's (1971) admonition that when essay questions are made too short and direct, they may lose their special value, and are then better replaced by multiple-choice questions. The trick is to make the question as direct as possible while still requiring that in answering it the test-taker exhibit the appropriate thinking process rather than merely filling in the "blank" (Coffman, 1971 p. 105).

2. Write questions that cover as wide a range of the material to be tested as possible (Jones, 1994).

If an essay test contains only a few questions that cover a narrow range of the content covered in class, it will lack validity. How well a student performs will then depend on whether the questions you ask represent the work he or she best understands. That will introduce an unnecessary bias (Dietrich, 1967).

3. Try your questions out on someone before you use them (Jones, 1994).

Show them to a colleague or fellow teacher, and try them out on some students if you can, maybe in a class review. See if the point that you are trying to test in the question is clear, or if it is open to misinterpretation (Jones, 1994 p. 145).

4. Develop a checklist of specific points to be counted in scoring each essay question (Jones, 1994).

This will increase the reliability of scoring. Tuckman (1988) calls this criterion point scoring, and describes it in some detail. In doing this, however, be careful not to eliminate from consideration any of the important factors that the question was designed to test (Coffman, 1971). Essay questions and their answers must be structured enough to be reliable measures of student capability, but not so structured as to lose their validity.

5. Attempt to construct a model answer for each question, one that incorporates all the specific points identified in step 4 (Jones 1994).

In preparing this answer, keep in mind the amount of time that students are likely to have to do it on the actual test. As an expert, the teacher should be able to come up with this answer in less time than the student is likely to have available.

6. Instead of scoring all the questions in a test, one after another, score the first question across the entire set of tests before proceeding to the second question (Tuckman 1991). This will make it easier to consistently score the answer to each question. Then, mix up the order of the tests and score all the second questions. Each time, cover the name of the student whose answer is being scored. Scoring this way helps minimize what has been called the "halo effect", the tendency to generalize judgment from one performance sample to another (Tuckman 1991).

7. Consider taking on a scoring partner or using a technique known as cooperative scoring (Diederich, 1967).

Find a fellow teacher to share scoring responsibility. Pick two random sets of tests and have each teacher score them. Then compare and discuss the results. This kind of

comparison and feedback can help both teachers overcome errors of judgment. Another idea would be to try grading the same set of tests more than once to see if the same results are found each time (Diederich, 1967).

8. Consider allowing students, on occasion, to complete their essay tests at home, without time limits, and with resource material available to them (Cooke, 1994).

This will eliminate reading and writing speed as a factor in test performance, and will, in some sense, make the test more authentic. Of course, it may introduce extraneous factors such as some other person contributing to the work.

9. Do not allow students to choose which essay question they want to answer unless questions of equal difficulty are chosen for the test (Cooke, 1994).

Evidence has shown that the more able students choose different items to answer than the less able students (Wiseman and Wrigley, 1958).

10. Build tests that include both multiple choice questions and essay questions (Tuckman, 1993).

This will enable the strengths to be maximized while the weaknesses are minimized. The strengths and weaknesses of the tester will be averaged at the same time as that of the test taker.

Concern about Traditional Testing

Much concern is being voiced about the validity of multiple-choice testing, and its critics are suggesting a variety of alternatives in its place (Bracey 1993). All alternatives proposed have been questioned in the areas of being "authentic", "direct", and measuring what they are supposed to. The most logical choice for today's High School level classroom seems to be the essay format (Ashburn, 1938). This type of testing is

regarded as being more complex, integrated, and more challenging than multiple-choice testing. It also focuses more on thinking skills and less on rote learning (Shepard, 1991).

The study of the issue of scoring reliability of authentic tests and the systematic development of techniques to overcome it have been largely, if not exclusively restricted to large-scale published tests, primarily used to measure writing skills. However, the majorities of the tests used by teachers in schools and on which assessments of student achievement are based are ones that teachers build and score themselves. Therefore, it is reasonable to ask if teachers can score authentic types of assessment with sufficient reliability to justify their use (Tuckman, 1993).

First, it may be useful to consider what an authentic assessment actually is. It is an actual sample of performance that requires that the test-taker construct a response, presumably a "realistic" one, that will be qualitatively judged against some performance standard or standards (Murchan, 1989, p. 64). Many assessments in athletics, for example, are of the authentic type. The competitive performances of divers, synchronized swimmers, figure skaters, and gymnasts are all of the authentic variety, and are evaluated by judges who are expert in judging each sport by virtue of training or experience. Despite their expertise, however, multiple judges are used and extreme judgments often discarded in recognition of the matter of scoring reliability. In boxing, as well, three judges are used, yet their judgments are often at variance with one another. In all popular team sports such as baseball, football, basketball, soccer, and hockey, judges are used to judge various aspects of performance such as illegal behaviors. Anyone who has ever witnessed any of these sports, especially with an emotional interest in the outcome, can testify to the concern with the reliability of scoring judgement (Tuckman, 1993).

Mr. Tuckman is trying to make a comparison here between judges in sporting events and teachers who judge their student's work. Probably many teachers would agree that this method would be a very accurate way of evaluating students, but at the same time, those same teachers would probably also say that time here would be the major drawback. After all, how many classroom teachers have the time to help their colleagues with grading?⁵

Unfortunately, there are few systematic studies of the scoring reliability of essay test responses by teachers, but one done by Asburn, (1938), among college professors, is quite revealing. Seventy-five history students were asked to compare and contrast life in a medieval castle and life in a medieval town (in 400 words) in terms of security, sanitation, job training, and recreation, while 75 English students were asked in the same number of words, to show that *The Canterbury Tales* affords a good picture of medieval society, and to analyze some of its characters. Three professors in each discipline independently read and scored the essays for that discipline. The level of scoring agreement did not exceed 50%. In fact, 40% of the students' score varied from passing to failing or vice versa depending on which professor did the scoring (Ashburn, 1938).

Even in the scoring of writing sample essays by trained scorers, specific kinds of scoring errors have been found to persist. One of these is context errors where essay ratings go up or down depending on the quality of the essays that are read immediately before and immediately after a given essay (Hughes and Keeling, 1984). Another source of errors is in the area of grammar, sentence structure, and spelling (Wolcott, 1988). Unfortunately, scoring judgments of what is written are affected by how it is written. The

grade of the student then would be determined partially by how much emphasis the particular teacher would be placing on these different areas (Hills, 1991).

Teachers have been urged to use essay items to measure higher mental processes (rather than knowledge acquisition), to use scoring keys or model responses to guide judgment, to obscure names of students, and to score each response twice (Tuckman, 1988). How likely is it that teachers (even college professors) follow these rules? There are so many biases such as fatigue, difficulty of scoring, and time restraints that probably most teachers don't (Wolcott, 1988).

We seem to be facing a real dilemma, according to Tuckman. What kind of testing should we recommend that teachers use, especially considering that teacher testing constitutes the main basis for student evaluation (Tuckman, 1991)? Should we recommend multiple-choice testing even though we are told (Shepard, and others, 1991) that it reinforces rote learning and (Hills, 1991) that teachers know relatively little about how to write good items? Or, should we recommend essay testing even though the scores teachers give may be as strong a reflection of their prior estimate of a student's capability as of the words written on the test paper (Coffman, 1968)? Would it be better to be potentially victimized by the item writer or by the item scorer? Well, as Tuckman (1993) says, "the item writer and the item scorer are the same person: the teacher"! The best solution is the one offered by Hills (1991), namely to train teachers in the area of testing. They themselves know that they are deficient (Carter, 1984).

Regardless of the type of item teachers use, the result will suffer in either validity, reliability, or both because teachers receive so little training in test construction (Carter, 1984). So while on the national level, critics of testing are clamoring for new and more

authentic testing formats, in the classroom where the "real" testing goes on, the issue is a much more fundamental one: how can we make better testers out of teachers. The answer seems obvious. Train them (Hills, 1991 p. 554)!

CHAPTER THREE

Description of Population

The groups used in the study were two high-school aged Biology II classes. The same teacher taught both classes during the same semester. The students used in the study were all juniors and seniors who chose to take Biology II as an elective. Since this is an advanced course, scores are expected to be high on both testing methods. Scores are also expected to be more consistent than if the study was performed on a general class. Students in this course are traditionally concerned with their grades, scientifically oriented, and are in the top ten percent of their class academically. Also, since it is a class that the students chose to take, it can be assumed there will be more effort to do well on their part. It can be assumed also that they are going to be doing their best on both types of testing. The sample population from which the study was constructed came from approximately one thousand students. This school is located in Fairlea, W.V. that is a rural community. The school contains students from basically six small communities, that all feed in to the high school. The communities are Alderson, Lewisburg, Renick, Ronceverte, White Sulphur Springs, and Williamsburg, all located geographically in southern West Virginia. The economic background of the students was mostly middle class. The occupation of their parents would range from such professional positions as doctors, lawyers, engineers and teachers to other trades such as coal minors, loggers, maintenance and other skilled positions.

Hypothesis

H₀: Students will achieve at the same level regardless of the style of test they are given.

H₁: When the students are taking the essay tests, they will achieve at higher levels than when they are taking multiple-choice tests.

Procedure

The data were obtained by using a combination of tests. The tests were generated from evaluation manuals sent with the text books which contain both multiple-choice and essay questions, along with questions made up by the classroom teacher over the years, also both multiple-choice and essay. Each test used in the research was constructed from only one of these two sources mentioned at a time. This helped to ensure that each test was put together using basically the same level of difficulty.

At the beginning of the course there was a discussion on test taking methods. These dealt with both multiple-choice and essay tests in both classes. The discussion dealt with how the students should prepare for each test, such as, notetaking, reading, laboratory investigations, reports, discussions, etc. The methods for reviewing for a test were also discussed. These methods can both be found in the appendix section of the thesis.

Each class knew ahead of time which type of test they would be taking for that grading period. Both classes were given the same format for dealing with each type of test and for how to review for the test. By doing this, it can be established that each student at least had in their possession the criterion for how to most effectively take a multiple-choice or essay test and also how to best review for each type of test. This helped establish consistency in the students as far as what is expected from them and in what they should expect to look for on the tests.

The length of the study was determined to be nine weeks since that corresponds to one grading period. Half of the grades for class 1 (2nd period) were determined totally by multiple-choice evaluation while the other half were determined by essay evaluation. Class 2 (5th period) was set up using the same procedure. At the end of the first 4 ½ week period, the evaluation methods for each half of each class were switched so that all students were given the opportunity to take part in both testing formats. Both classes were being tested on the same material at the same time so that difficulty of text was not a variable. The Biology II course used in the study had a prerequisite of an A or B in Biology I, therefore all students used in the study were good students in science. Furthermore, when the students were placed in each of the two classes, they were picked randomly by the computer to determine which section (1 or 2) they would be in. It can be assumed then, that there was a good mixture of "A" and "B" students in each section.

Research Question

Does the testing style affect the achievement level of students?

CHAPTER FOUR

Statistical Analysis

The research question answered by this study was, whether there is a difference in test scores when comparing multiple choice tests and essay tests to the achievement level in students. The study was started at the start of the second grading period, therefore the grades from the first grading period were used as a pretest comparison. The study period consisted of a combination of both testing styles. All multiple-choice test scores have been totaled and averaged for each student. The same has been done for the essay tests.

Comparisons have been made between each testing method, both for each individual and for each class, by using a two-sample t-test. Since the two classes were selected randomly it can be assumed that they are relatively equal academically, therefore this is a valid comparison. One thing to consider is that since each class was tested on the same material at the same time with half of each class taking multiple-choice and the other half essay, the material covered could more easily conform to one or the other styles of testing. The mean score for method A (essay), and method B (multiple-choice), S^2_{pooled} , Est. σ^2_{A-B} and Est. σ_{A-B} were calculated and listed in table 4-1.

Table 4-1

Method A Essay	Method B Multiple Choice	S^2_{pooled}	Est σ_{A-B}	Est σ_{A-B}
Mean 83	Mean 78	100.63	4.03	2.007

Descriptive Statistics

The mean score for method A (essay) was 83%. The mean score for method B (multiple-choice) was 78%. The S^2_{pooled} value when considering the two methods was

calculated to be 100.63. The estimated σ^2_{A-B} was found to be 4.03 and the estimated σ_{A-B} was 2.007.

T-tests

A two-sample t-test was used to analyze the mean scores of the multiple-choice and essay tests to determine if there was a significant difference in achievement.

Table 4-2

t-value	critical t	v	α
2.49	2.0	48	.05

The results showed a t-value of 2.49, critical t was 2.0 with 48 degrees of freedom at the .05 level of significance. The two methods were determined not to be equal in evaluating the achievement levels of students. This was due to the fact that $2.49 > 2.0$ at the .05 level of significance. The null hypothesis that was tested was: students will achieve at the same level regardless of the style of test they are given. Based on the results of the t-test, the null hypothesis is rejected and the alternate hypothesis is accepted. The alternate hypothesis is: when the students are taking the essay tests, they will achieve at higher levels than when they are taking multiple-choice tests.

CHAPTER FIVE

Summary, Conclusions and Recommendations

Summary

The purpose of this study was to compare two different testing methods and see how they relate to retention of information. The two methods used were:

1. Essay
2. Multiple-choice

In place of a pre-test the previous nine week average was used for comparison with the two testing methods. This grading period consisted of a combination of both testing styles. At the conclusion of the study, the grades of the two testing styles were compared by the use of a two sample t-test. At the .05 level of significance there was a noticable difference in the achievement levels of the two groups.

Conclusions

The study showed that students taking essay tests seem to do better on the average than those taking multiple-choice tests. Essay tests, according to this study, are a better way of assessing the retention levels of students than multiple choice tests.

When compared to their 1st nine weeks average, which was an average that came from a variety of testing methods, of the 50 students, 15 improved their average on the multiple choice tests, 7 students remained exactly the same and 28 scored higher on the essay tests.

Further study on this comparison of essay versus multiple-choice testing styles needs to be done, such as:

1. Extend the length of the study to include the entire school year instead of only one

grading period

2. Use a larger sample size.
3. Ensure that the original grade that was used as a pretest was composed of an equal combination of essay and multiple choice tests.

REFERENCE PAGE

- Jones, Russell, w, "Performance and Alternative Assessment Techniques": U.S. Massachusetts, 1994, ppg. 142-150.
- Bernice Jenson Bragsted and Sharyn Mueller Stumpf, Study Skills and Motivation, (1987), ppg. 309-312.
- Elissa Sommerfield and Frances Bailey, Study Skills Guide: A Survivor's Manuel, Texas: Educational Skills, 1983), ppg. 29-40.
- Tuckman, Bruce W, "The Essay Test: A Look at the Advantages and Disadvantages", NASSP Bulletin, (Oct 1993), ppg. 20-26.
- Takata, Susan R, "The Guided Essay Examination for Sociology and other Courses", Teaching Sociology, (April 1994), ppg. 189-194.
- Green, Kinsey B, "Test Item Construction in the Cognitive Domain", (1979), ppg. 52-55.
- White, Edward M. "An Apologia for the Timed Impromptu Essay Test", College Composition and Communication, (Feb 1995) ppg. 30-45.
- House, Ernest R, and Others, "Definition of Content in Social Studies Testing: (1987) ppg. 143-160.
- McDaniel, Mark A and Others, "Test Expectancy, Study Strategies and Recall of Prose", (1991) ppg. 143-150.
- Tuckman, Bruce W. "Evaluating the Alternative to Multiple Choice Testing for Teachers", Contemporary Education, (1991) ppg. 299-300.
- Molek, Carol and Goss, Barbara, "The Parent/Student Study Skills Connection", (1990) ppg. 141-152.

- Thiede, Keith W. "Recognition Versus Recall Test Formats: A Correlational Analysis", (1991) ppg. 32-36.
- Murchan, Damien P. "Essay Versus Objective Achievement Testing in the Context of Large-Scale Assessment Programs", (1989) ppg. 61-65.
- Abplanalp, Paul. "Application of Latent Image Technology to the Multiple Choice Test Format", Journal of Optometric Education, (1995) ppg. 118-123.
- Cooke, Brian P. "Rethinking Teaching and Testing: Quality in the Classroom", (1994) ppg. 120-127.
- Haigh, William. "Using the Computer to Solve Problems by the Guess and Test Method", School Science and Mathematics, (Feb 1993) ppg. 92-95.
- Powell, Zen-Hsiu, Emily. "The psychological Impacts of Computerized Adaptive Testing", Educational Technology, (Oct 1994) ppg. 41-47.
- Kolstead, Rosemarie; Kolstead, Robert, "Applications of Conventional and Non-Restrictive Multiple Choice Examination Items", Clearing House, (Jul-Aug 1994), ppg. 317-319.
- Bracey, Gerald W. "Assessing the New Assessments", Principal, (Jan 1993), ppg. 34-36.
- Shavelson, Richard J. and Baxter, Gayle P. "What We've Learned About Assessing Hands on Science", Educational Leadership, (May 1992) ppg. 20-25.
- Ashburn R.R. "An Experiment in the Essay-Type Question", Journal of Experimental Education, (1938) ppg. 1-3.
- Carter, K. "Do Teachers Understand Principles for Writing Tests?" Journal of Teacher Education, (1984) ppg. 57-60.

- Coffman, W.E. Curfman, D.A. "A Comparison of Two Methods of Reading Essay Examinations", American Educational Research Journal, (1968) ppg. 99-107.
- Dietrich, P.B. "Cooperative Preparation and Rating of Essay Tests", English Journal, (1967) ppg. 573-584.
- Tuckman, B.W. Testing for Teachers, 2nd Ed. San Diego: Harcourt Brace Jovanovich, 1988. (Evaluating the Alternative to Multiple Choice Testing for Teachers), Contemporary Education (1991) ppg. 299-300.
- Hills, J.R. "Apathy Concerning Grading and Testing", Phi Delta Kappan, (1991) ppg. 540-555.
- Hushes, D.C. and Keeling, B. "Use of Model Essays to Reduce Context Effects in Essay Scoring" Journal of Educational Measurement, (1984) ppg. 277-281.
- Wolcott, W. "Discrepancies in Essay Scoring", Resources in Education, (1988)
- Shepard, L. "Interview on Assessment Issues", Educational Researcher, (1991) ppg. 21-23.